



(51) International Patent Classification:
G06V 30/19 (2022.01) *G06V 30/418* (2022.01)
G06V 30/412 (2022.01)

PETROV, Petar Marinov; c/o BLUE PRISM LIMITED,
2 Cinnamon Park, Crab Lane, Fearnhead, Warrington WA2
OXF (GB).

(21) International Application Number:
PCT/EP2023/065079

(74) Agent: **HUSBAND, Benjamin** et al.; Carpmaels & Rans-
ford LLP, One Southampton Row, London WC1B 5HA
(GB).

(22) International Filing Date:
06 June 2023 (06.06.2023)

(81) Designated States (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG,
KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA,
NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO,
RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,
TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
22177366,6 06 June 2022 (06.06.2022) EP

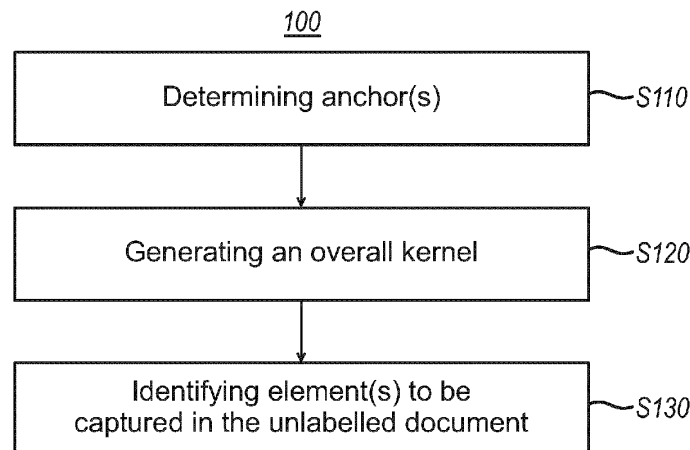
(71) Applicant: **BLUE PRISM LIMITED** [GB/GB]; 2 Cinna-
mon Park, Crab Lane, Fearnhead, Warrington WA2 OXP
(GB).

(72) Inventors: **REID, John Edward**; c/o BLUE PRISM
LIMITED, 2 Cinnamon Park, Crab Lane, Fearnhead,
Warrington WA2 OXP (GB). **AROCA-OUELLETTE,
Miguel André**; c/o BLUE PRISM LIMITED, 2 Cinnamon
Park, Crab Lane, Fearnhead, Warrington WA2 OXP (GB).

(84) Designated States (*unless otherwise indicated, for every
kind of regional protection available*): ARIPO (BW, CV,
GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST,

(54) Title: METHODS AND SYSTEMS FOR PERFORMING DATA CAPTURE

FIG. 1



(57) Abstract: There is provided a method (100) for performing data capture in an unlabelled document based on a labelled document, wherein the labelled document contains one or more elements each labelled with respect to a field. The method comprises: determining (S110) one or more anchors, wherein an anchor comprises a pair of points that match between the labelled document and the unlabelled document; generating (S120) an overall kernel using the determined one or more anchors; and identifying (S130), using the overall kernel in a kernel machine, one or more elements to be captured in the unlabelled document, wherein the one or more elements to be captured correspond in field to the one or more labelled elements contained in the labelled document.

SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*